**Storage Exercises**

1. ABC Corporation is starting on a Capex project to set up a big data cluster to host their data warehouse. They have a current data warehouse of size 150TB. The daily data rate is 200 GB and is supposed to increase by 10% each year. They are currently using tape backup but want to eliminate that. Estimate the number of machines required for the cluster.

Hint: Update the numbers in the template to get the number of machines.

Answer: Approximately 93 machines.

1. XYZ Health Analytics is setting up a big data cluster to host their data analytics platform. They want to load 200TB of archived data and also load daily data of 300GB. The data is supposed to grow at 15% per year. They want to pack 10 machines into a rack. Estimate the number of machines and number of racks required.

Answer: Approximately 130 machines and 13 racks.

1. ABC Telecom is setting up a big data cluster to host their data analytics platform. They want to load 500 TB of archived data. They have 10 channel partners. They get 200 GB of data from each of their channel partners every month. They want to plan the cluster for 3 years with 20% growth per year. Estimate the number of machines required.

Hint: The cluster sizing template is for 1 year and uses 365 days. Modify it to use for 3 years with the 12 months in a year.

Answer: Approximately 242 machines. with 28.8TB for 1st year, 34.56TB for second year and 41.472TB for third year plus the archived data.

1. An enterprise wants to take advantage of real-time big data technology for their data analytics. They have data centers in two cities and want to ensure data is replicated across these centers. They want to set up real time data processing using a No-SQL database. They have data coming in at 100GB per day at each datacenter. Estimate the number of machines needed.

Approximately 24 machines in each data center (with replication of 4 : 3 in local and 1 in remote datacenter)

1. A global financial enterprise has offices in three continents, North America, Europe and Asia. They want to set up real time data processing using a No-SQL database. They have 500TB of data in North America, 200TB in Europe and 100TB in Asia. They want a proper fault-tolerant setup. Suggest a proper big data setup with number of machines in each continent.

Hint: You may suggest replication of 3 within content and an extra copy outside the continent. Choose the continent for replication based on proximity.

Answer: NA : 200 machines, Eu: 160 machines, Asia: 67 machines

Per machine storage: For the above activities, pick a per machine setup of 10TB storage.